



PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10187754 A**(43) Date of publication of application: **21 . 07 . 98**

(51) Int. Cl.

G06F 17/30(21) Application number: **08356219**(22) Date of filing: **25 . 12 . 96**(71) Applicant: **NEC CORP**(72) Inventor: **RI KÔ
YAMANISHI KENJI**(54) **DEVICE AND METHOD FOR CLASSIFYING
DOCUMENT**

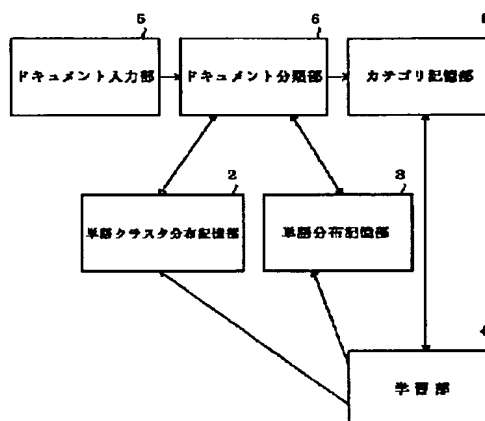
logarithmic likelihood is classified.

COPYRIGHT: (C)1998,JPO

(57) Abstract:

PROBLEM TO BE SOLVED: To make high-accuracy document classification realizable, by making the distribution of word clusters in a category correspond to the linear link model of distribution of words in respective word clusters.

SOLUTION: A document classifying part 6 receives a document inputted from a document input part 5. Then, the distribution of word clusters in respective categories stored in a word cluster distribution storage part 2 is referred to, the distribution of words in respective word clusters stored in a word distribution storage part 3 is referred to, and the distribution of word clusters in each category and the linear link model of distribution of words in the respective word clusters are made correspondent to that category and defined as inputted document data. The negative logarithmic likelihood of linear line model corresponding to each category is calculated for these data and the document inputted to the category corresponding to the linear link modal having the minimum calculated negative



THIS PAGE BLANK (USPTO)

(19) 日本国特許庁 (J P)

(12) 特 許 公 報 (B 2)

(11) 特許番号

第2940501号

(45) 発行日 平成11年(1999) 8月25日

(24) 登録日 平成11年(1999) 6月18日

(51) Int.Cl.⁶

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/401

3 1 0 D

請求項の数 2 (全 11 頁)

(21) 出願番号 特願平8-356219

(22) 出願日 平成 8 年(1996) 12月25日

(65) 公開番号 特開平10-187754

(43) 公開日 平成10年(1998) 7月21日

審査請求日 平成 8 年(1996) 12月25日

(73) 特許権者 000004237

日本電気株式会社

東京都港区芝五丁目 7 番 1 号

(72) 発明者 李 航

東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

(72) 発明者 山西 健司

東京都港区芝五丁目 7 番 1 号 日本電気株式会社内

(74) 代理人 弁理士 加藤 朝道

審査官 平井 誠

最終頁に続く

(54) 【発明の名称】 ドキュメント分類装置及び方法

1

(57) 【特許請求の範囲】

【請求項 1】 カテゴリと該カテゴリに分類されたドキュメントを記憶するカテゴリ記憶部と、

カテゴリにおける単語クラスタの分布を記憶する単語クラスタ分布記憶部と、

単語クラスタにおける単語の分布を記憶する単語分布記憶部と、

(a) 前記カテゴリ記憶部に記憶される、カテゴリと、該カテゴリに分類されたドキュメントと、を参照して、各カテゴリに対応する単語クラスタを作成し、

(b) 前記各カテゴリについて、該カテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、前記各単語クラスタにおける単語の分布を推定し、

(c) 推定された前記各単語クラスタにおける単語の分

2

布を、前記単語分布記憶部に記憶し、

(d) さらに各カテゴリにおける単語クラスタの分布を推定し、推定された前記各カテゴリにおける単語クラスタの分布を、前記単語クラスタ記憶部に記憶する学習手段と、

新たに入力されるドキュメントをドキュメント分類部に格納するドキュメント入力手段と、

(e) 前記ドキュメント入力手段から入力されたドキュメントを受け取り、

10 (f) 前記単語クラスタ分布記憶部に記憶される各カテゴリにおける単語クラスタの分布を参照し、及び、前記単語分布記憶部に記憶される各単語クラスタにおける単語の分布を参照し、各カテゴリに、該カテゴリにおける単語クラスタの分布と、各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されたドキュメ

ントをデータとみなし、該データに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、

(g) 計算された負対数尤度の最も小さい線形結合モデルの対応するカテゴリに入力されたドキュメントを分類するドキュメント分類手段と、

を備えることを特徴とするドキュメント分類装置。

【請求項2】 (a) カテゴリと該カテゴリに分類されたドキュメントを記憶するカテゴリ記憶部に記憶されるカテゴリと該カテゴリに分類されたドキュメントを参照し、各カテゴリに対応する単語クラスタを作成し、

(b) 各カテゴリについて、各単語クラスタにおける単語の分布を推定し、推定した単語の分布を第1の記憶領域に記憶し、

(c) さらに、各カテゴリにおける単語クラスタの分布を推定し、推定された単語クラスタの分布を第2の記憶領域に記憶しておく、

(d) 新たに入力されたドキュメントを受け取った際には、単語クラスタの分布を記憶する前記第2の記憶領域から各カテゴリにおける単語クラスタの分布を参照すると共に、単語の分布を記憶する前記第1の記憶領域から各単語クラスタにおける単語の分布を参照し、

(e) 各カテゴリについて、該カテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されたドキュメントをデータとみなし、該データに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、

(f) 負対数尤度の最も小さい線形結合モデルの対応するカテゴリに入力されたドキュメントを分類する、ことを特徴とするドキュメント分類方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、インターネットのホームページの自動分類、電子図書館における文献検索、特許出願情報の検索、電子化された新聞記事の自動分類、マルチメディア情報の自動分類等の用途に適用して好適とされる情報の分類や検索技術に関する。

【0002】

【従来の技術】情報の分類や検索の分野において、ドキュメント分類（「文章分類」、「テキスト分類」ともいう）装置の開発は大きな課題である。ここで、「ドキュメント分類」とは、予め幾つかのカテゴリを設けておき、さらに一部のドキュメントがそれぞれどのカテゴリに属するかを判断し、該当するカテゴリに、ドキュメントを分類し、システムに記憶した後、システムは記憶された情報から知識を自動的に獲得し、これ以降、獲得できた知識を基に、新たに入力されたドキュメントを自動的に分類する、ことをいう。

【0003】従来、幾つかのドキュメント分類装置が提案されている。その中でも、Salton（サルトン）らの提案するドキュメント分類装置がよく知られている。例

えば文献⁽¹⁾（G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", New York: McGraw Hill, 1983）が参照される。この装置は、ドキュメントに現れる単語の頻度ベクトルとカテゴリにおける単語の頻度ベクトルの間のコサイン値をドキュメントとカテゴリ間の距離とみなし、距離の最も小さいカテゴリにドキュメントを分類する、ものである。

【0004】また、Guthrie（グスリー）らの提案するドキュメント分類装置は、単語をクラスタにまとめるものとして、注目されている。例えば文献⁽²⁾（Guthrie Louise, Walker Elbert, and Guthrie Joe, "Document Classification by Machine: Theory and Practice," Proceedings of the 15th international Conference on Computational Linguistics (COLING'94), page 1059-1063, 1994）が参照される。

【0005】図15は、上記Guthrieらが提案するドキュメント装置の構成を示す図である。図15を参照すると、このドキュメント装置は、ドキュメント入力部505、ドキュメント分類部503、単語クラスタ分布記憶部502、カテゴリ記憶部501、及び、学習部504を備えて構成され、ドキュメントに現れる単語（あるいは、キーワード、タームともいう）を幾つかの単語クラスタに分類し、単語クラスタの出現分布を基にドキュメントの分類を行うものである。

【0006】Guthrieらの提案するドキュメントの分類装置は、単語をクラスタに分類しているため、Saltonらの提案するドキュメント分類装置よりも精度の高い分類ができる。

【0007】以下では、簡単な例を通じて、Guthrieらの提案するドキュメント分類装置について説明する。予め「野球」と「サッカー」という2つのカテゴリを設けるとする。なお、これは利用者が設定する。

【0008】そして、幾つかのドキュメントについて、この2つのカテゴリのどちらに属するかを判断し、該当するカテゴリに分類した後、これらの情報をカテゴリ記憶部501に記憶したとする。図8は、2つのカテゴリのドキュメントに現れる単語の出現度数の一例である。

【0009】Guthrieらの提案するドキュメント装置では、学習部504は、カテゴリ「野球」に対して、単語クラスタ「野球」を作成し、「サッカー」に対して、単語クラスタ「サッカー」を作成する。カテゴリ「サッカー」のドキュメントに現れず、カテゴリ「野球」のドキュメントに度数を1回以上（一般的にはN回以上）に現れた単語を単語クラスタ「野球」に分類し、一方、カテゴリ「野球」のドキュメントに現れず、カテゴリ「サッカー」のドキュメントに度数を1回以上（一般的にはN回以上）に現れた単語を単語クラスタ「サッカー」に分類する。さらに、残りの単語を、単語クラスタ「その

他」に分類する。

【0010】すると、図9に示すような、3つの単語クラスタが得られる。すなわち、図8の各カテゴリに現れる単語出現頻度情報から、クラスタ「野球」には、単語「ベース」、及び「投手」が分類され、クラスタ「サッカー」には、「ゴール」が分類され、クラスタ「その他」には、「試合」、及び「観衆」が分類される。

【0011】また、図10に示すような、2つのカテゴリのドキュメントに現れる単語クラスタ（クラスタ野球、クラスタサッカー、クラスタその他）の出現頻度も 10

$$P(X=x) = (f(X=x) + 0.5) / (F + 0.5 * k) \quad \dots(1)$$

【0014】但し、 $P(X=x)$ は x の起きる確率で、 $f(X=x)$ は F 回の観測結果の中の x の起きる回数である。また k は X のとる値の種類の数である。

【0015】図11は、カテゴリ「野球」とカテゴリ「サッカー」における、単語クラスタ（クラスタ野球、クラスタサッカー、クラスタその他）の分布を示したものである。

【0016】ドキュメント分類では、ドキュメント分類部503は、ドキュメント入力部505から新しいドキュメントの入力を受け、単語クラスタ分布記憶部502に記憶される各カテゴリにおける単語クラスタの分布を参照し、入力されたドキュメントをデータとみなし、そのデータが各カテゴリにおける単語クラスタの分布から生起される確率を計算し、生起確率の最も大きい分布に対応するカテゴリに、入力されたドキュメントを分類する。具体的には、以下のような処理を行う。

【0017】ドキュメント分類部503は、図12に示すような入力（観衆、投手、ベース、ベース、ゴール）

log確率（データ | カテゴリ「野球」）

$$\begin{aligned} &= \log 0.52 + \log 0.43 + \log 0.43 + \log 0.43 + \log 0.05 \\ &= -8.92 \end{aligned}$$

log確率（データ | カテゴリ「サッカー」）

$$\begin{aligned} &= \log 0.58 + \log 0.05 + \log 0.05 + \log 0.05 + \log 0.37 \\ &= -15.19 \end{aligned}$$

【0020】カテゴリ「野球」からの生起確率の方が、カテゴリ「サッカー」からの生起確率よりも大きいので、入力されるドキュメントを、カテゴリ「野球」に分類する。

【0021】

【発明が解決しようとする課題】しかしながら、上記したGuthrieらの提案になるドキュメント分類装置は、以下記載の3つの問題点を有している。

【0022】（1）第1の問題点は、同じ単語クラスタに分類された単語が同等に扱われる、ということである。

【0023】例えば、「ベース」と「投手」が同じく単語クラスタ「野球」に分類され、そのどちらかが現れ 50

得られる。

【0012】学習部504は、次に、各カテゴリに、そのカテゴリにおける単語クラスタの分布を対応させ、Laplace（ラプラス）推定量を用いて、単語クラスタの分布を推定し、得られる単語クラスタの分布を単語クラスタ分布記憶部502に記憶する。Laplace推定量を用いた確率パラメータの推定式を次式（1）に示す。

【0013】

【数1】

を受けるとする。ドキュメント分類部503は、入力されたドキュメントに現れる単語を、その単語が属する単語クラスタによって置き換え、図13に示すようなデータを作成する。すなわち、観衆、投手、ベース、ベース、ゴールは、それぞれクラスタその他、クラスタ野球、クラスタ野球、クラスタ野球、クラスタサッカーに置き換えられる。

【0018】ドキュメント分類部503は、次に、単語クラスタ分布記憶部502から、図11に示すカテゴリ「野球」とカテゴリ「サッカー」における単語クラスタの分布を参照し、図13のデータは、ある単語クラスタの分布から生成されるとし、そのデータが、図11に示すカテゴリ「野球」とカテゴリ「サッカー」における単語クラスタの分布から生起される確率を、以下のように計算する。但し、ここでは、取り扱いやすいように、確率値の対数をとっている。

【0019】

【数2】

ば、単語クラスタ「野球」が現れるとしている。しかし、「ベース」のドキュメントにおける出現度数が、「投手」の出現度数よりも多く、新しいドキュメントに「ベース」が現れた場合、そのドキュメントに「投手」が現れる場合に比べて、より高い精度と確信度で、ドキュメントをカテゴリ「野球」に分類できるはずである。しかしながら、上記したGuthrieらの装置では、このようなことはできない。

【0024】（2）第2の問題点は、単語クラスタを作成する時の単語出現度数の閾値の設定が困難である、ということである。

【0025】上記したGuthrieらの提案する装置では、カテゴリ「サッカー」のドキュメントに現れず、カテゴ

り「野球」のドキュメントにN回以上現れた単語を単語クラスタ「野球」に分類し、カテゴリ「野球」のドキュメントに現れず、カテゴリ「サッカー」のドキュメントにN回以上現れた単語を単語クラスタ「サッカー」に分類している。そして、それ以外の単語を単語クラスタ「その他」に分類している。

【0026】この場合、Nの設定が大きな問題となる。すなわちNの値が大きければ、クラスタ「野球」とクラスタ「サッカー」にそれぞれ分類される単語が減り、クラスタ「その他」に分類される単語が増えることになる。その結果、入力されたドキュメントが、どのカテゴリに属するか判断できない場合が増える。

【0027】一方、Nの値が小さければ（例えば、N=1）、クラスタ「野球」とクラスタ「サッカー」に分類される単語が増える。しかし、1回しか現れない単語も何回も現れる単語も同じように取り扱われることから、分類の精度が下がる。

【0028】(3)第3の問題点は、複数のカテゴリのドキュメントに現れるが、全体としては、あるカテゴリのドキュメントに偏って現れる単語を有効に利用することができない、ということである。

【0029】例えば、カテゴリ「野球」とカテゴリ「サッカー」のドキュメントに現れる単語とその出現度数が、図14に示すようなものであるとする。図14を参照すると、「ゴール」は主にカテゴリ「サッカー」のドキュメントに現れるが、カテゴリ「野球」のドキュメントにも現れている。

【0030】上記Guthrieらの提案する装置では、この場合、「ゴール」を単語クラスタ「その他」に分類してしまい、単語「ゴール」のよく現れるドキュメントをカテゴリ「サッカー」に分類する、ことはできない。

【0031】したがって、本発明は、上記事情に鑑みてなされたものであって、その目的は、単語がある確率で単語クラスタに属するとし、各カテゴリにそのカテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させることにより、上記した従来のドキュメント装置の問題点を全て解消し、高精度のドキュメント分類を実現可能とするドキュメント分類装置を提供することにある。

【0032】

【課題を解決するための手段】前記目的を達成するため、本発明のドキュメント分類装置は、まず、単語を単語クラスタに分類する時、該単語がある確率でその単語クラスタに属するとし、さらに、各カテゴリに、そのカテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させる。

【0033】より詳細には、本発明のドキュメント分類装置は、カテゴリと該カテゴリに分類されたドキュメントを記憶するカテゴリ記憶部と、カテゴリにおける単語クラスタの分布を記憶する単語クラスタ分布記憶部と、

単語クラスタにおける単語の分布を記憶する単語分布記憶部と、(a)前記カテゴリ記憶部に記憶される、カテゴリと、該カテゴリに分類されたドキュメントと、を参照して、各カテゴリに対応する単語クラスタを作成し、

(b)前記各カテゴリについて、該カテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、前記各単語クラスタにおける単語の分布を推定し、(c)推定された前記各単語クラスタにおける単語の分布を、前記単語分布記憶部に記憶し、(d)さらに各カテゴリにおける単語クラスタの分布を推定し、推定された前記各カテゴリにおける単語クラスタの分布を、前記単語クラスタ記憶部に記憶する学習手段と、新たに入力されるドキュメントをドキュメント分類部に格納するドキュメント入力手段と、

(e)前記ドキュメント入力手段から入力されたドキュメントを受け取り、(f)前記単語クラスタ分布記憶部に記憶される各カテゴリにおける単語クラスタの分布を参照し、及び、前記単語分布記憶部に記憶される各単語クラスタにおける単語の分布を参照し、各カテゴリに、該カテゴリにおける単語クラスタの分布と、各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されたドキュメントをデータとみなし、該データに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、(g)計算された負対数尤度の最も小さい線形結合モデルの対応するカテゴリに入力されたドキュメントを分類するドキュメント分類手段と、を備えることを特徴とする。

【0034】

【発明の実施の形態】本発明の実施の形態について以下に説明する。本発明は、その好ましい実施において、カテゴリと該カテゴリに分類されたドキュメントを記憶するカテゴリ記憶部(図1の1)と、カテゴリにおける単語クラスタの分布を記憶する単語クラス分布記憶部(図1の2)と、単語クラスタにおける単語の分布を記憶する単語分布記憶部(図1の3)と、学習部(図1の4)と、新たに入力されるドキュメントをドキュメント分類部(図1の6)に格納するドキュメント入力部(図1の5)と、及びドキュメント分類部(図1の6)を備えて構成される。

【0035】本発明の実施の形態において、学習部(図1の4)は、その処理フローの一例を示した図2を参照すると、(a)カテゴリ記憶部(図1の1)に記憶される、カテゴリと、該カテゴリに分類されたドキュメントと、を参照して、各カテゴリに対応する単語クラスタを作成し(図2のステップ101、102)、(b)各カテゴリにおける単語クラスタの分布を推定し、推定された前記各カテゴリにおける単語クラスタの分布を、前記単語クラスタ記憶部に記憶し(図2のステップ103)、(c)前記各カテゴリに、該カテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布

の線形結合モデルを対応させ、前記各単語クラスタにおける単語の分布を推定し、推定された前記各単語クラスタにおける単語の分布を、単語分布記憶部（図1の3）に記憶する（図2のステップ104）。

【0036】本発明の実施の形態において、ドキュメント分類部（図1の6）は、その処理フローの一例を示した図7を参照すると、（a）ドキュメント入力部（図1の5）から入力されたドキュメントを受け取り（図7のステップ301）、（b）単語クラスタ分布記憶部（図1の2）に記憶される各カテゴリにおける単語クラスタの分布を参照し、及び、単語分布記憶部（図1の3）に記憶される各単語クラスタにおける単語の分布を参照し（図7のステップ302）、（c）各カテゴリに、そのカテゴリにおける単語クラスタの分布と、各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されたドキュメントをデータとみなし、該データに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、計算された負対数尤度の最も小さい線形結合モデルの対応するカテゴリに入力されたドキュメントを分類する（図7のステップ303）。

【0037】

【実施例】上記した本発明の実施の形態について更に詳細に説明すべく、本発明の実施例について図面を参照して以下に説明する。

【0038】図1は、本発明のドキュメント分類装置の第1の実施例の構成を示す図である。図1を参照すると、本実施例のドキュメント分類装置は、カテゴリ記憶部1、単語クラスタ分布記憶部2、単語分布記憶部3、学習部4、ドキュメント入力部5、及び、ドキュメント分類部6を備えて構成されている。

【0039】カテゴリ記憶部1は、カテゴリとそのカテゴリに分類されたドキュメントを記憶する。

【0040】学習部4は、カテゴリ記憶部1に記憶されるカテゴリとそのカテゴリに分類されたドキュメントを参照して、各カテゴリに対応する単語クラスタを作成し、各カテゴリに、そのカテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、各単語クラスタにおける単語の分布を推定し、推定できた各単語クラスタにおける単語の分布を単語の分布記憶部3に記憶し、さらに各カテゴリにおける単語クラスタの分布を推定し、推定できた各カテゴリにおける単語クラスタの分布を単語クラスタ記憶部2に記憶する。

【0041】ドキュメント入力部5は、新しいドキュメントを入力する。

【0042】ドキュメント分類部6は、ドキュメント入力部5から入力されたドキュメントを受け取り、単語クラスタ分布記憶部2に記憶される、各カテゴリにおける単語クラスタの分布、を参照し、また単語分布記憶部3に記憶される、各単語クラスタにおける単語の分布、を

参照し、各カテゴリに、そのカテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されたドキュメントをデータとみなし、そのデータに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、計算できた負対数尤度の最も小さい線形結合モデルの対応するカテゴリに入力されたドキュメントを分類する。

【0043】本実施例のドキュメント分類装置の処理動作について、図14に示す例に即して以下に説明する。

10 【0044】カテゴリ記憶部1は、カテゴリとそのカテゴリに分類されたドキュメントを記憶する。一般的には、カテゴリを c_1, c_2, \dots, c_n と表す。例えば、記憶されるカテゴリとそのカテゴリに分類されたドキュメントに現れる単語の出現度数が、図14に示すようなものであるとする。ここでは、カテゴリは、「野球」と「サッカー」の2つである。

【0045】学習の際、学習部4は、図2に示すフローチャートに従う処理を行う。

20 【0046】すなわち学習部4は、まず、カテゴリ記憶部1に記憶されるカテゴリと、そのカテゴリに分類されたドキュメントと、を参照し（ステップ101）、カテゴリに対応する単語クラスタを作成する（ステップ102）。具体的には、学習部4はカテゴリに1対1に単語のクラスタを作成する。

【0047】作成された単語クラスタを、 k_1, k_2, \dots, k_n と表す。図9に示す例では、単語クラスタ「野球」と単語クラスタ「サッカー」を作成する。

30 【0048】単語の各カテゴリのドキュメントにおける出現頻度をみて、あるカテゴリにおける出現頻度が40%以上である場合、そのカテゴリに対応する単語クラスタにその単語を分類する。このように分類できない単語を以降の処理で無視する。図14に示す例に対して、図3に示すような単語クラスタが得られる。

【0049】学習部4は、次に、各単語クラスタにおける単語の分布を推定し、推定できた各単語クラスタにおける単語の分布を単語の分布記憶部3に記憶する。一般的には、単語クラスタ k_i における単語の分布を、 $P(W | k_i)$ と表す。但し、 k_i はある単語クラスタを表し、確率変数 W は単語クラスタ k_i に属する単語を値とする。

【0050】学習部4は、以下のように、各単語クラスタにおける単語の分布を推定する。

【0051】単語クラスタ k_i における単語 w の出現確率 $P(w | k_i)$ （次式（2））に従って推定する。

【0052】 $P(w | k_i) = f(w) / F \dots (2)$

50 【0053】但し、 $P(w | k_i)$ は単語クラス k_i における単語 w の出現確率で、 $f(w)$ は単語 w のすべてのドキュメントにおける出現度数、 F は単語クラスタ k_i におけるすべての単語のすべてのドキュメントにおける出現度数である。

【0054】図3に示す単語クラスタにおける単語の分布は、図4に示すようなものとなる。

【0055】学習部4は、次に、各カテゴリに、そのカテゴリにおける単語クラスタ分布と各単語クラスタにおける単語の分布による線形結合モデルを対応させる。

$$P(W|c) = \sum_{i=1}^n P(W|k_i) \times P(k_i|c) \quad \dots(3)$$

$$P(W|k_i) = \begin{cases} P(W|k_i) & W \in k_i \\ 0 & \text{otherwise} \end{cases} \quad \dots(3-1)$$

【0058】学習部4は、各カテゴリにおける単語クラスタの分布を推定し、推定できた各カテゴリにおける単語クラスタの分布を単語クラスタ分布記憶部2に記憶する。一般的には、単語クラスタの分布を $P(K|c)$ と表す。但し、 c はあるカテゴリを表し、 K は単語クラスタを値とする。

【0059】学習部4は、具体的には、例えば、隠れ変数によるマルコフチェインモンテカルロ法を用いて、カ

$$P(k_i|c) = \theta_i, \quad P(W|k_i) = p_i(W) \quad \dots(4)$$

$$P(W|\theta) = \sum_{i=1}^n \theta_i \times p_i(W), \quad \theta = (\theta_1, \theta_2, \dots, \theta_n) \quad \dots(5)$$

【0062】次に、隠れ変数 Z を導入する。 Z は、長さが n で、1つの値が1でその他の値がすべて0であるようなベクトルを値とする。例えば、次式(6)は Z の取る値の例である。

$$Z = (0, \dots, 0, 1, 0, \dots, 0) \quad \dots(6)$$

【0064】次に、隠れ変数モデルを定義する。隠れ変

$$P(W, Z|\theta) = \sum_{i=1}^n \delta_i(Z) \times \theta_i \times p_i(W) \quad \dots(7)$$

ただし、

$$\delta_i(Z) = \begin{cases} 1 & \text{if } Z = (z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) = (0, \dots, 0, 1, 0, \dots, 0) \\ 0 & \text{otherwise} \end{cases} \quad \dots(8)$$

$$P(W|\theta) = \sum_{Z \in Z'} P(W, Z|\theta) \quad \dots(9)$$

【0066】いま、上式(7)における θ の推定を問題として考える。以下では、 Z と θ を繰り返し互いにランダムサンプリングすることによって θ の推定を行う。そのフローチャートを図5に示す。

【0067】まずDirichlet(ディリクレ)分布 $D(a_1, a_2, \dots, a_n; \theta)$ を定義する。ここで、「Dirich

【0056】線形結合モデルは、次式(3)のように定義される。

【0057】

【数3】

テゴリにおける単語クラスタの分布 $P(K|c)$ を推定する。

【0060】表記上簡単のため、以下 $P(k_i|c)$ と $P(W|k_i)$ を、次式(4)で表す。すると、上式(3)のモデルは、次式(5)のようになる。

【0061】

【数4】

数モデルは、 W と Z の同時分布を用いて、次式(7)、(8)のように表現する。そして W に関する周辺分布は、次式(9)となる。

【0065】

【数5】

40 let分布」とは、次式(10)に示す密度関数をもつ確率分布のことをいう。なお、 a_1, a_2, \dots, a_n はパラメータであり、 Γ はガンマ関数である。

【0068】

【数6】

$$D(a_1, a_2, \dots, a_n; \theta)$$

14

$$= \frac{\Gamma(a_1 + \dots + a_n)}{\Gamma(a_1)\Gamma(a_2)\dots\Gamma(a_n)} \theta_1^{a_1-1} \theta_2^{a_2-1} \dots \theta_n^{a_n-1}, \quad 0 \leq \theta_i \leq 1, \sum_{i=1}^n \theta_i = 1$$

... (10)

【0069】 θ の初期値を適当に求め、 $\theta^{(0)}$ とする
(ステップ201)。次にサンプリングを繰り返し、 θ
と Z を求めていく。 $\theta^{(1)}$ と $Z^{(i)}$ を1回目にサンプリン
グで得られる値とする。まず1+1回目の繰り返しサン
プリングでは、次式(11)の分布に従って、 $Z_i(i$

= 1, 2, ..., N)の値をランダムサンプリングする
(ステップ203)。

【0070】

【数7】

$$Z_i^{(1+1)} \sim P(Z_i | W_i, \theta^{(1)})$$

... (11)

【0071】 $W_i(i=1, 2, \dots, N)$ は観測データ
である。ここでは、 $P(Z_i | W_i, \theta)$ は、次式(1
2)で与えられる。

【0072】

【数8】

$$P(Z_i | W_i, \theta) = \frac{P(W_i, Z_i | \theta)}{P(W_i | \theta)}$$

... (12)

【0073】 $Z_i^{(1+1)}(i=1, 2, \dots, N)$ の値、す
なわち $(Z^N)^{(1+1)}$ が得られた後、次式(13)の分布
に従って、 $\theta^{(1+1)}$ の値をランダムサンプリングする

(ステップ204)。

【0074】

【数9】

$$\theta^{(1+1)} \sim P(\theta | W^N, (Z^N)^{(1+1)})$$

... (13)

【0075】事前分布を $D(a_1, a_2, \dots, a_n; \theta)$
とおくと、 $P(\theta | W^N, Z^N)$ は、次式(14)、(1
5)、(16)で与えられる。

【0076】

【数10】

$$P(\theta | W^N, Z^N) = D(a_1 + t_1, a_2 + t_2, \dots, a_n + t_n; \theta)$$

... (14)

$$Z^N = Z_1, Z_2, \dots, Z_N, Z_i = (z_{i1}, z_{i2}, \dots, z_{in})$$

... (15)

$$t_j = \sum_{i=1}^n z_{ij}$$

... (16)

【0077】上記のようにサンプリングを繰り返す。指
定された回数を実行したら、処理を終了する。 $\theta^{(1+1)}$
を推定結果とする。

【0078】このような推定法によって得られた θ は、
サンプリング回数が十分大きい場合、事後分布 $P(\theta |$
 $W^N)$ に従って、サンプリングされたものと近似的にみ
なすことができる。

【0079】上式(10)のサンプリングにおいて、実
際には、パラメータ θ の取り得る値を制限する。具体的

$$\theta_j = \frac{1 - \theta_i}{n - 1}, \quad j \neq i$$

... (17)

【0082】学習部4は、上記のように、各カテゴリに
おける単語クラスタの分布を推定し、推定できた各カテ
ゴリにおける単語クラスタの分布を単語クラスタ記憶部
2に記憶する。

【0083】推定できたカテゴリ「野球」とカテゴリ

「サッカー」における単語クラスタの分布の一例を、図

には、パラメータ空間の量子化を行い、以下のように θ
の取り得る値を決める。カテゴリが c_i である時、 θ_i の
値を0.5から1までの r (例えば、 $r=0.05$)刻
みの値をとるとする。

【0080】こうして θ_i が決まった後、残りのパラメ
ータは、次式(17)のように決める。

【0081】

【数11】

6に示す。

【0084】ドキュメント分類の際、ドキュメント入力
部5は、ドキュメント分類部6に、新しいドキュメント
を入力する。

【0085】ドキュメント分類部6は、ドキュメント入
力部5から入力されるドキュメントを受け取り、単語ク

ラスタ分布記憶部2に記憶される各カテゴリにおける単語クラスタ分布を参照し、単語分布記憶部3に記憶される各単語クラスタにおける単語分布を参照し、各カテゴリに、そのカテゴリにおける単語クラスタ分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させ、入力されるドキュメントをデータとみなし、そのデータに対する、各カテゴリの対応する線形結合モデルの負対数尤度を計算し、計算できた負対数尤度のもっとも

$$L(d | \text{カテゴリ「野球」})$$

$$= -\log 0.25 \cdot 0.90 - \log 0.75 \cdot 0.90 - \log 0.75 \cdot 0.90 - \log 0.80 \cdot 0.10 \\ = 6.93$$

$$L(d | \text{カテゴリ「サッカー」})$$

$$= -\log 0.25 \cdot 0.05 - \log 0.75 \cdot 0.05 - \log 0.75 \cdot 0.05 - \log 0.80 \cdot 0.95 \\ = 16.19$$

【0089】但し、入力されたテキストは、図12に示したものとし、各カテゴリにおける単語クラスタの分布は図6、各単語クラスタにおける単語の分布は図4にそれぞれ示すようなものであるとする。

【0090】負対数尤度の最も小さいカテゴリにドキュメントdを分類する。この場合、カテゴリ「野球」による負対数尤度が小さいので、ドキュメントをカテゴリ「野球」に分類する。

【0091】次に、本発明のドキュメント分類装置の第2の実施例について説明する。本発明の第2の実施例の構成は、図1に示したものと同様とされ、カテゴリ記憶部1、単語クラスタ分布記憶部2、単語分布記憶部3、学習部4、ドキュメント入力部5、及びドキュメント分類部6を備える。

【0092】本発明の第2の実施例のドキュメント分類

$$L(\theta) = \sum_{j=1}^N \log \left[\sum_{i=1}^M \theta_i P_i(W_j) \right] \quad \dots(18)$$

$$\nabla L(\theta) = \left[\frac{\partial L(\theta)}{\partial \theta_1} \dots \frac{\partial L(\theta)}{\partial \theta_M} \right] \quad \dots(19)$$

$$\theta_i^{(l+1)} = \frac{\theta_i^{(l)} \exp(\eta \nabla L(\theta^{(l)})_i)}{\sum_{i=1}^M \theta_i^{(l)} \exp(\eta \nabla L(\theta^{(l)})_i)} \quad \dots(20)$$

$$\theta_i^{(l+1)} = \theta_i^{(l)} (\eta \nabla L(\theta^{(l)})_i - 1) + 1 \quad \dots(21)$$

【0096】

【発明の効果】以上説明したように、本発明のドキュメント分類装置においては、単語がある確率で単語クラスタに属するとし、さらに、各カテゴリに、そのカテゴリにおける単語クラスタの分布と各単語クラスタにおける

小さいカテゴリに入力されたドキュメントを分類する。

【0086】図7は、ドキュメント分類部の処理を説明するためのフローチャートである。

【0087】入力されたドキュメントd（データ）に対する、カテゴリcに対応する線形結合モデルの負対数尤度 $L(d | c)$ を、以下のように計算する。

【0088】

【数12】

装置の学習部4は、前記第1の実施例の装置の学習部と、異なるアルゴリズムで、各カテゴリにおける単語クラスタの分布を推定する。本発明の第2の実施例のドキュメント分類装置のこれ以外の部分は、前記第1の実施例の装置と同じである。以下では、相違点のみ説明する。

【0093】本実施例では、各カテゴリにおける単語クラスタの分布の推定問題を、次式(18)を最大にする問題、すなわち、最尤推定の問題として考える。

【0094】次式(20)、(21)のいずれかの式に従って、繰り返し計算することにより、 θ を求める。なおlは繰り返し計算のインデックス(index)であるとする。また、 $\eta > 1$ は係数であるとする。

【0095】

【数13】

単語の分布の線形結合モデルを対応させている。本発明によれば、このような構成としたことにより、従来Guthrieらの提案するドキュメント分類装置よりも高精度のドキュメント分類を実現することができる。

【0097】また、本発明においては、単語がある確率

で単語クラスタに属するとしているので、同じ単語クラスタに分類された単語が同等に扱われるという上記従来技術の問題点を解決することができる。さらに、単語クラスタを作成する時の単語出現度数の閾値の設定が困難であるという、従来技術の問題点も解消することができる。

【0098】そして、本発明においては、各カテゴリに、そのカテゴリにおける単語クラスタの分布と各単語クラスタにおける単語の分布の線形結合モデルを対応させることによって、複数のカテゴリのドキュメントに現れるが、全体としてはあるカテゴリのドキュメントに偏って現れる単語を有効に利用できないという、従来技術の問題点をも解消することができる。

【図面の簡単な説明】

【図1】本発明のドキュメント分類装置の一実施例の構成を示す図である。

【図2】本発明の第1の実施例の学習部の処理を説明するためのフローチャートである。

【図3】本発明の第1の実施例を説明するための図であり、単語クラスタとそれに属する単語を示す図である。

【図4】本発明の第1の実施例を説明するための図であり、単語クラスタにおける単語の分布を示す図である。

【図5】本発明の第1の実施例の学習部の推定処理を説明するためのフローチャートである。

【図6】各カテゴリにおける単語クラスタの分布を示す図である。

【図12】

【図13】

【図3】

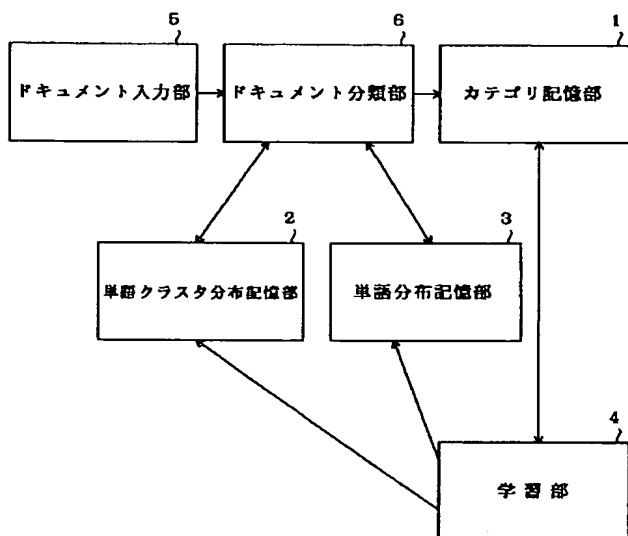
観衆、投手、ベース、ベース、ゴールクラスタその他、クラスタ野球、クラスタ野球、クラスタ野球、クラスタサッカー クラスタ野球：

ベース、投手

クラスタサッカー：

ゴール、キック

【図1】



【図4】

【図9】

クラスタ野球：

ベース：0.75、投手：0.25

クラスタサッカー：

ゴール：0.80、キック：0.20

クラスタ野球：

ベース、投手

クラスタサッカー：

ゴール

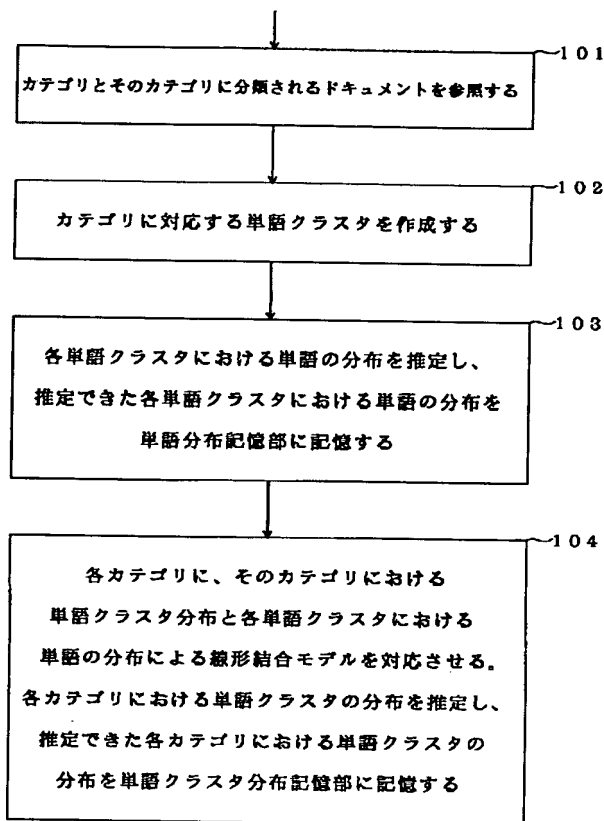
クラスタその他：

試合、観衆

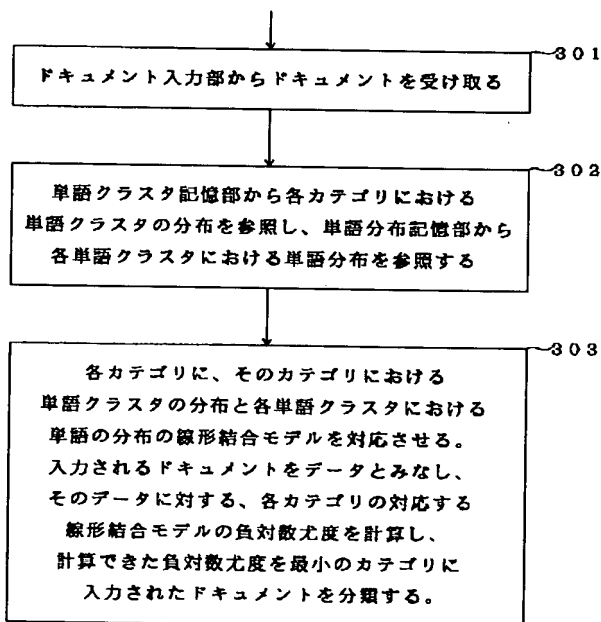
【図6】

カテゴリ\クラスタ	クラスタ野球	クラスタサッカー
野球	0.90	0.10
サッカー	0.05	0.95

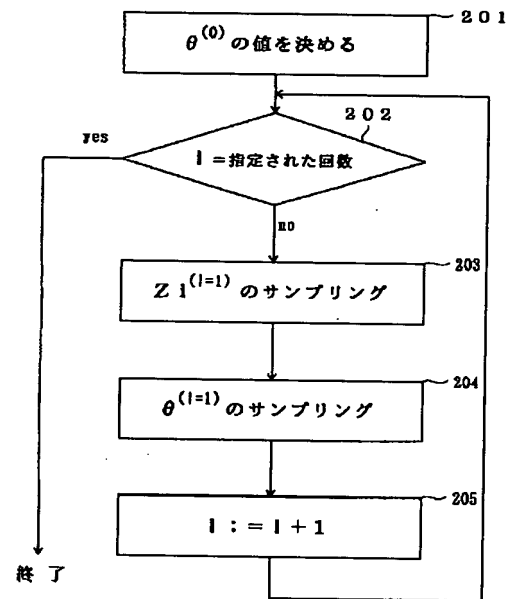
【図2】



【図7】



【図5】



【図8】

カテゴリ\単語	ベース	投手	ゴール	試合	観衆
野球	3	1	0	3	2
サッカー	0	0	3	3	2

【図10】

カテゴリ\クラスタ	クラスタ野球	クラスタサッカー	クラスタその他
野球	4	0	5
サッカー	0	3	5

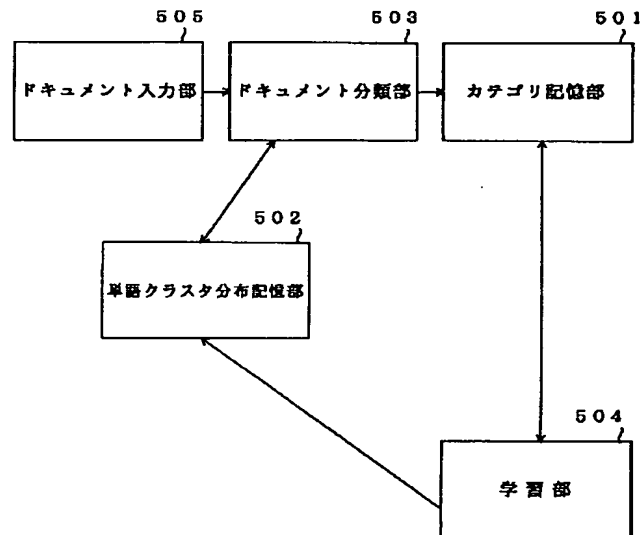
【図11】

カテゴリ\クラスタ	クラスタ野球	クラスタサッカー	クラスタその他
野球	0.43	0.05	0.52
サッカー	0.05	0.37	0.58

【図14】

カテゴリ\単語	ベース	投手	ゴール	キック	観衆
野球	3	1	1	0	2
サッカー	0	0	3	1	2

【図15】



フロントページの続き

(56) 参考文献 特開 平8-287097 (JP, A)
 岩山真, 徳永健伸, 「自動文書分類の
 ための新しい確立モデル」, 情報処理学
 会研究報告 Vol. 94, No. 37 (94-
 FI-33), pp 47-52 (平成6年5月
 18日)

(58) 調査した分野(Int.Cl.⁶, DB名)
 G06F 17/30

THIS PAGE BLANK (USPTO)